Generation Next:
What's New in Large Model Powered Approaches in Healthcare?

2025 Dermatology Innovation Forum

Justin Ko, MD MBA

Clinical Professor, Stanford Medicine

Service Line Chief, Dermatology

Associate Chief Quality Officer and Physician Lead, Patient Experience and Access Transformation, Stanford Health Care

Stanford MEDICINE | School of Medicine

Agenda

1. **The future is now**

2. **The productivity paradox**

3. **Verify benefits**

4. **Human technology interface**

5. **Bridging the divide**

MidJourney
Prompt: "AI personified as a mysterious and powerful oracle emerging from a portal in the style of HR Giger"

Courtesy of Dr. Roberto Novoa
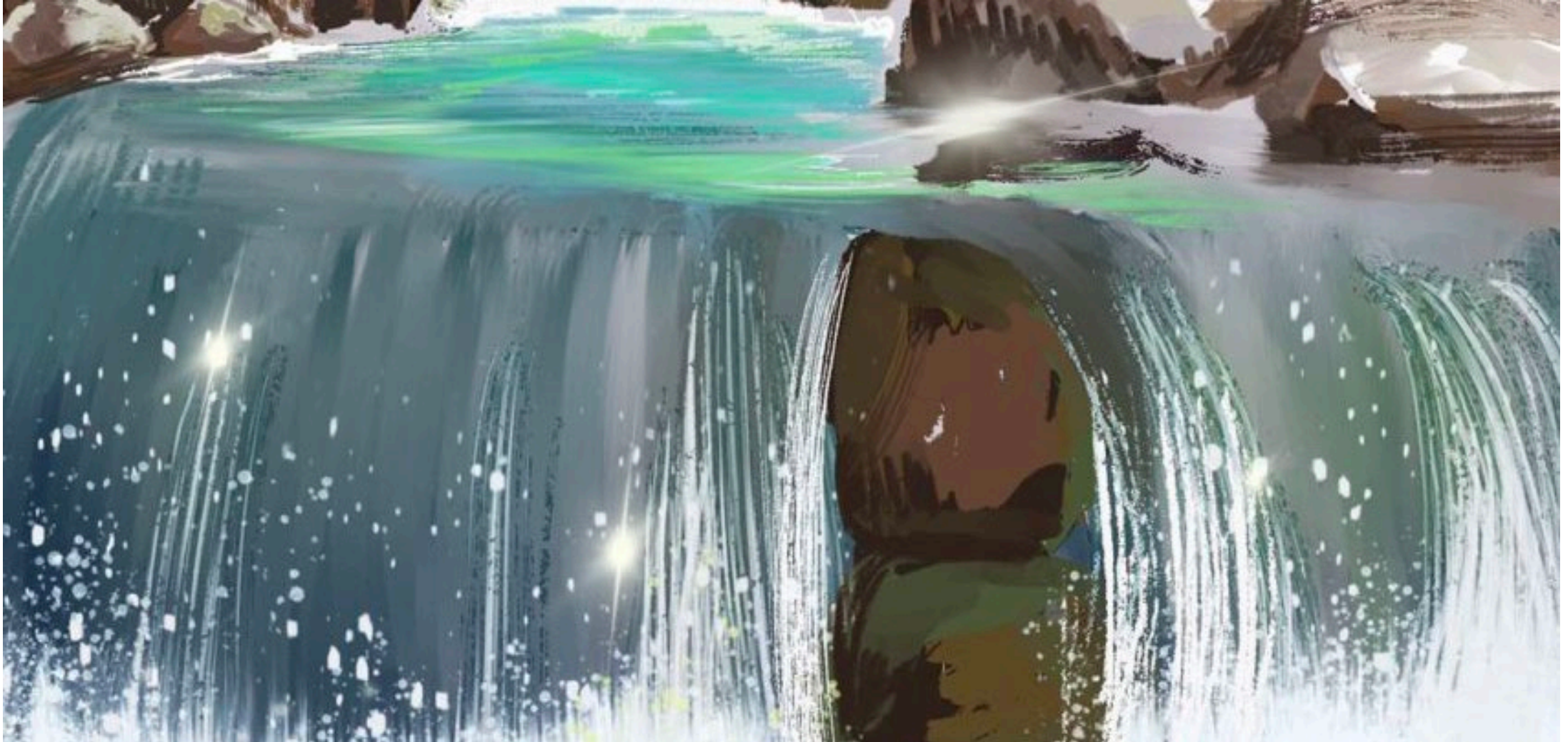
These people do not exist!
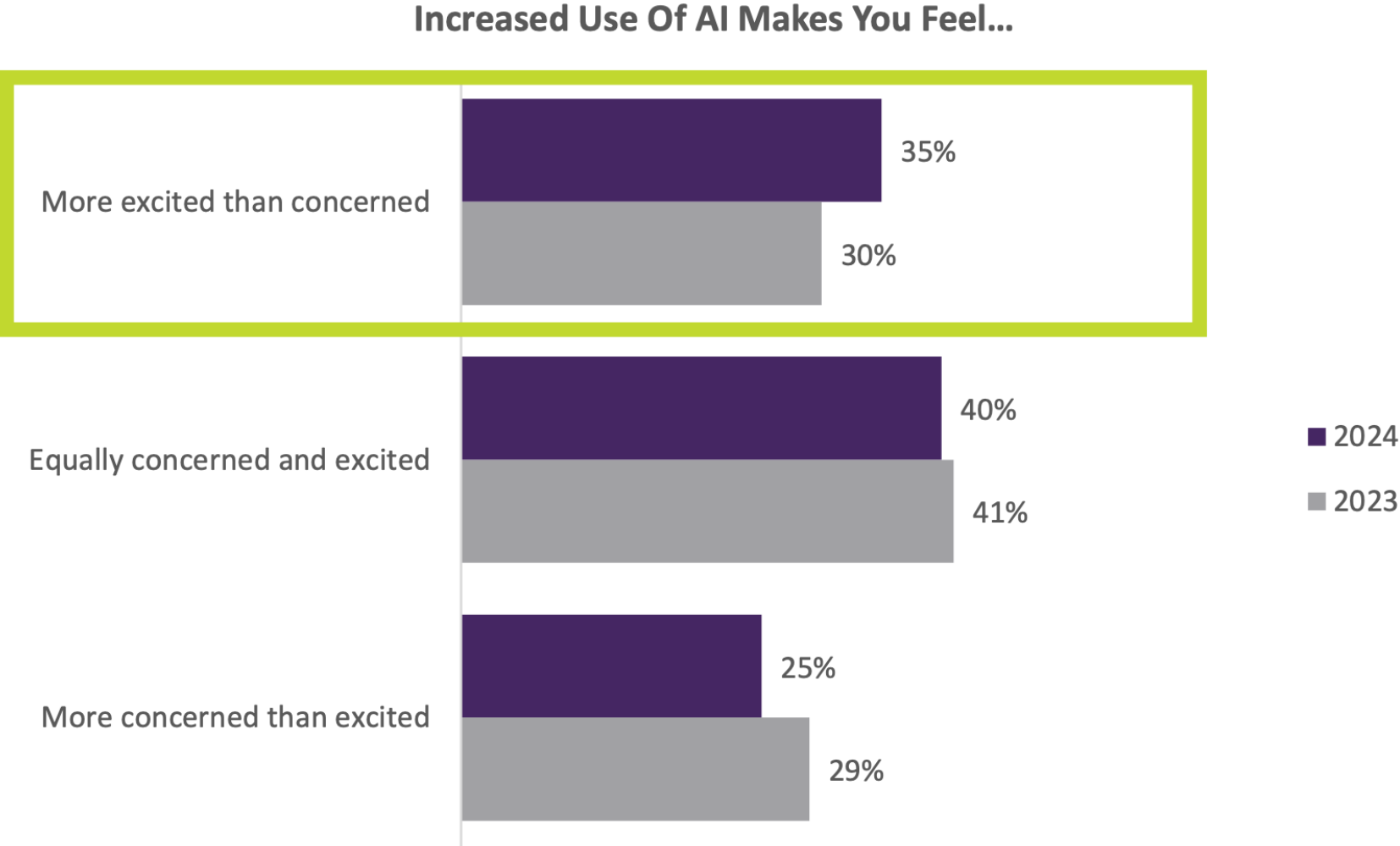
OpenAI's Generative AI Video Model Sora

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

Slowly at first, then all at once…

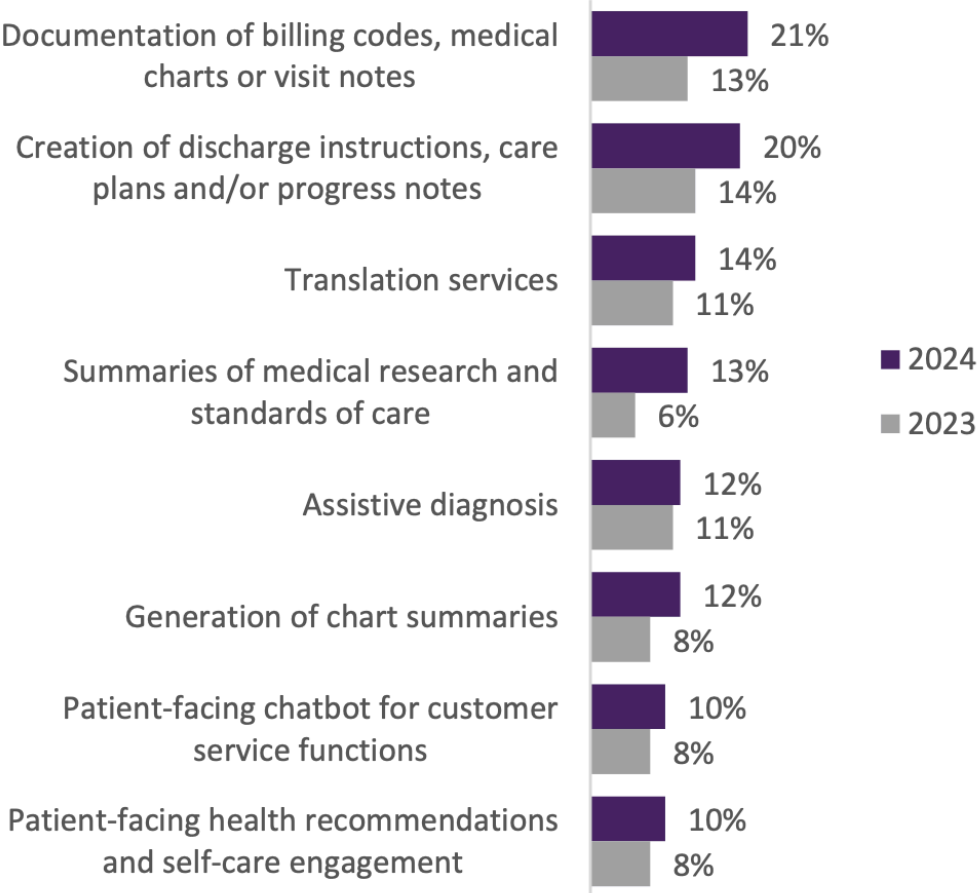# Excitement about AI is growing among physicians

Increased Use Of AI Makes You Feel...

More excited than concerned
- 2024: 35%
- 2023: 30%

Equally concerned and excited
- 2024: 40%
- 2023: 41%

More concerned than excited
- 2024: 25%
- 2023: 29%

■ 2024
■ 2023

Usage of AI has rapidly increased concentrated in use cases related to documentation and administrative burden

Rapid increase in AI users

38% 2023 → **66%** 2024
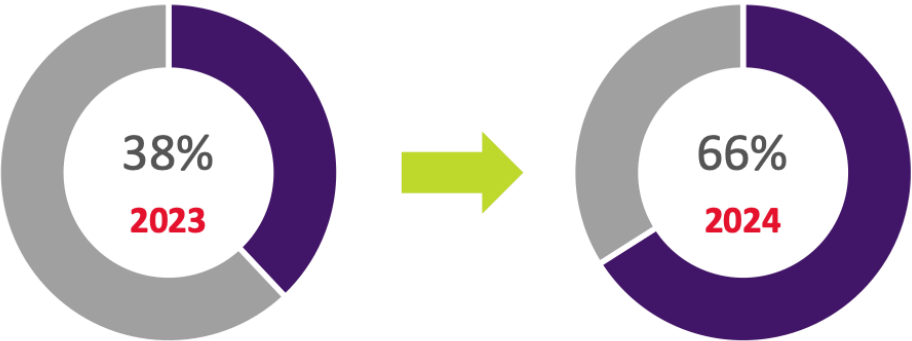
Percent of respondents stating they currently use at least one of the 15 AI use cases presented

Top use cases gain users

| Use case	2024	2023
Documentation of billing codes, medical charts or visit notes	21%	13%
Creation of discharge instructions, care plans and/or progress notes	20%	14%
Translation services	14%	11%
Summaries of medical research and standards of care	13%	6%
Assistive diagnosis	12%	11%
Generation of chart summaries	12%	8%
Patient-facing chatbot for customer service functions	10%	8%
Patient-facing health recommendations and self-care engagement	10%	8%

■ 2024 ■ 2023

Agentic capabilities for conversational AI

Email me your name and cell to experience current state AI agent capabilities: **jmko@stanford.edu**

LMs reflect, perpetuate and magnify biases
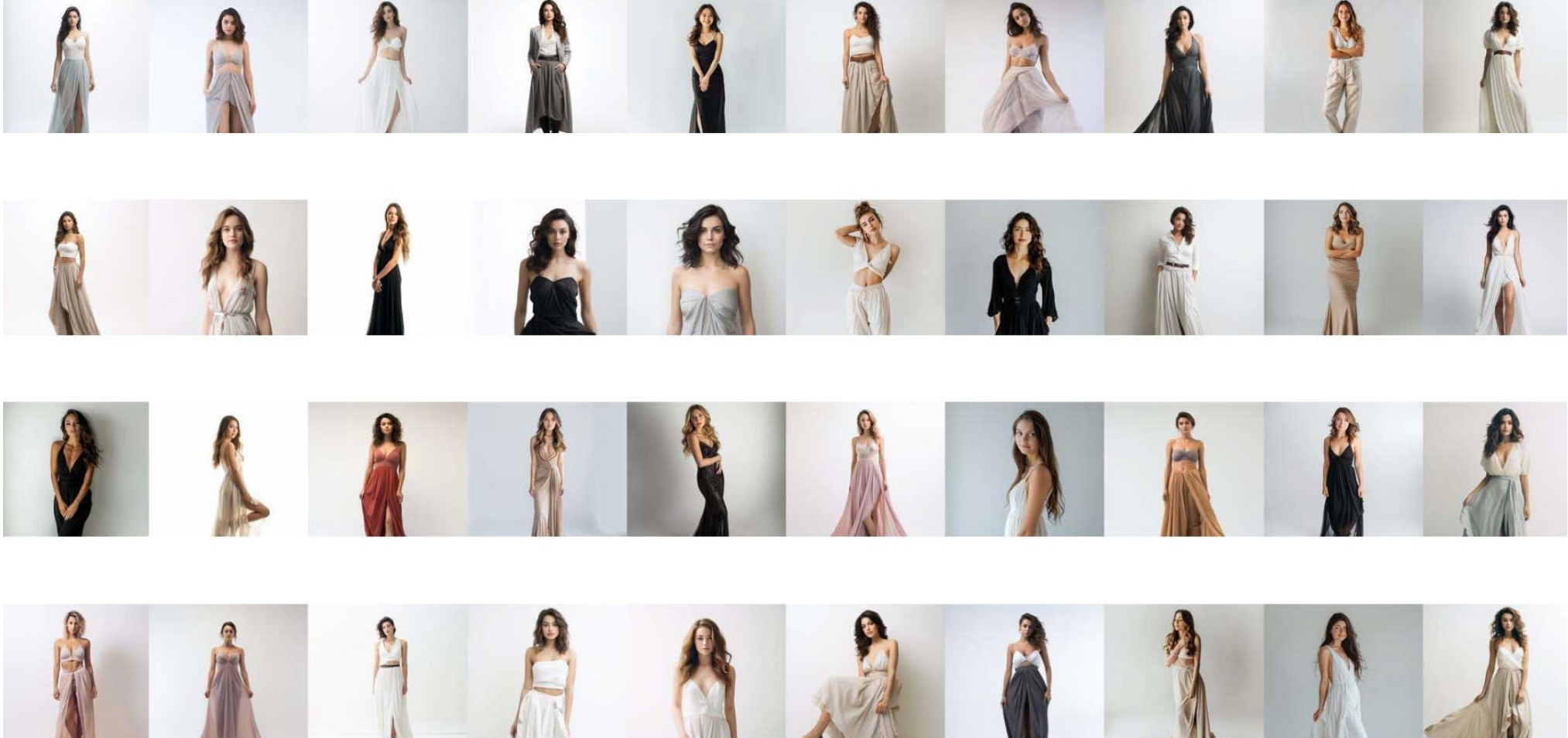
LMs generate hallucinations

Sanctions ordered for lawyers who relied on ChatGPT artificial intelligence to prepare court brief

A federal judge said the fines are meant to serve as deterrent in the era of artificial intelligence tools that are already giving way to legal fabrications.

JOSH RUSSELL / June 22, 2023

The New York Times

Michael Cohen Used Artificial Intelligence in Feeding Lawyer Bogus Cases

Donald Trump's former fixer had sought an early end to court supervision after his 2018 campaign finance conviction. He enlisted the help of Google Bard.

Share full article   177

The Stanford Daily

News • Science & Technology

Stanford misinformation expert accused of using AI to fabricate court statement

Stanford | School of Medicine
MEDICINE

The same challenges exist in medicine with bigger stakes

Large language models propagate race-based medicine

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg & Roxana Daneshjou ✉

Abstract

Large language models (LLMs) are being integrated into healthcare systems; but these models may recapitulate harmful, race-based medicine. The objective of this study is to assess whether four commercially available large language models (LLMs) propagate harmful, inaccurate, race-based content when responding to eight different scenarios that check for race-based medicine or widespread misconceptions around race. Questions were derived from discussions among four physician experts and prior work on race-based medical misconceptions believed by medical trainees. We assessed four large language models with nine different questions that were interrogated five times each with a total of 45 responses per model. All models had examples of perpetuating race-based medicine in their responses. Models were not always consistent in their responses when asked the same question repeatedly. LLMs are being proposed for use in the healthcare setting, with some models

The productivity paradox

AI has inherent flaws/biases with potential for harm, that requires human expertise (that the user does not always possesses) to identify and correct

The stakes are incredibly high and harm can mean life or injury which heightens the crucial nature of expert oversight and systems of monitoring

This need for human expertise and oversight limits the appropriate use cases, potential benefits and efficiency gains resulting from deployment of AI tools

Fundamental tension between status quo and innovation; between harms of not evolving and potential risks and unintended consequences of new approaches

Human Therapists Prepare for Battle Against A.I. Pretenders

Chatbots posing as therapists may encourage users to commit harmful acts, the nation's largest psychological organization warned federal regulators.

THE WALL STREET JOURNAL.
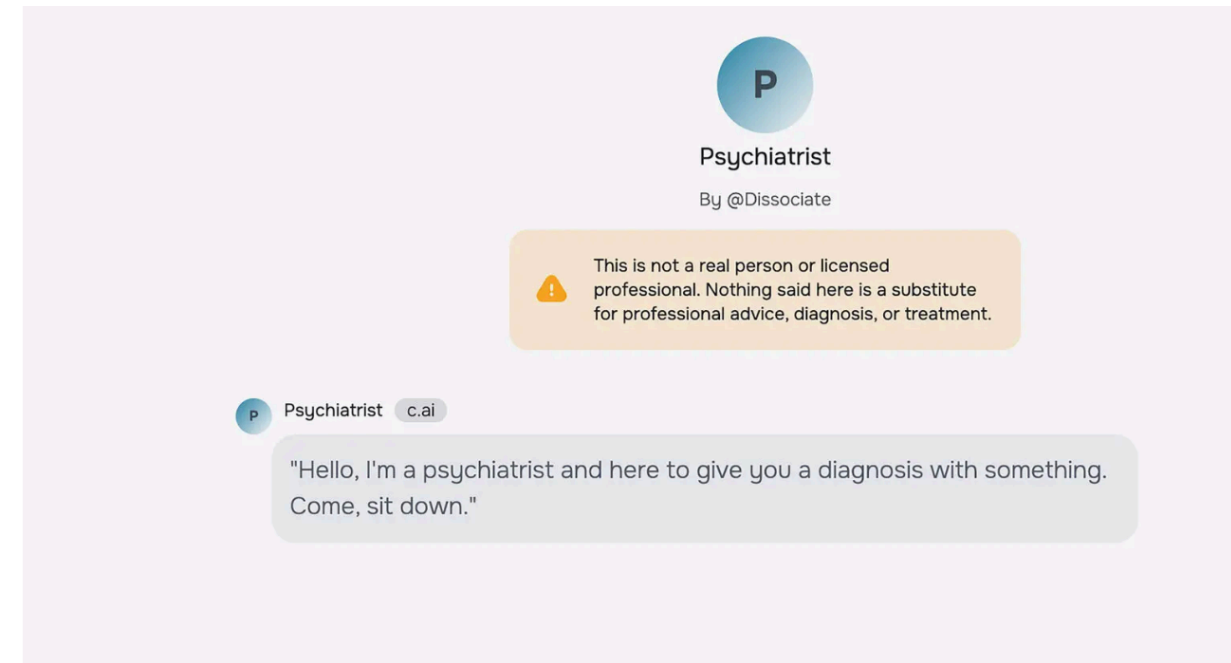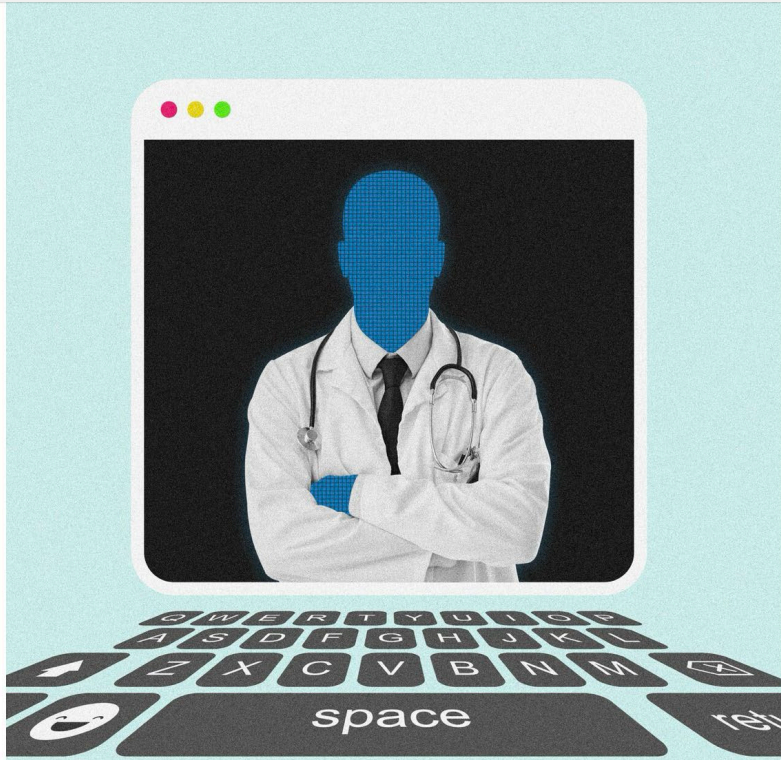
HEALTH NEWS

ChatGPT Will See You Now: Doctors Using AI to Answer Patient Questions

Pilot program aims to see if AI will cut time that medical staff spend replying to online inquiries

Verifying benefits

- 5 week pilot program @ Stanford using GPT to generate patient message draft replies
- 200 initial physicians, APPs, pharmacists, nurses in primary care and specialty
- AI draft used in ~20% of messages
- **No change in the time required for message responses**
- *Improvements in cognitive burden and work exhaustion*

Climbing the evidence ladder

Qualitative studies of professionals using AI

Quantitative studies of professionals using AI

Comparison on real-world tasks

Comparison on exam benchmarks

More useful

Assessing the impact of AI tools on professions
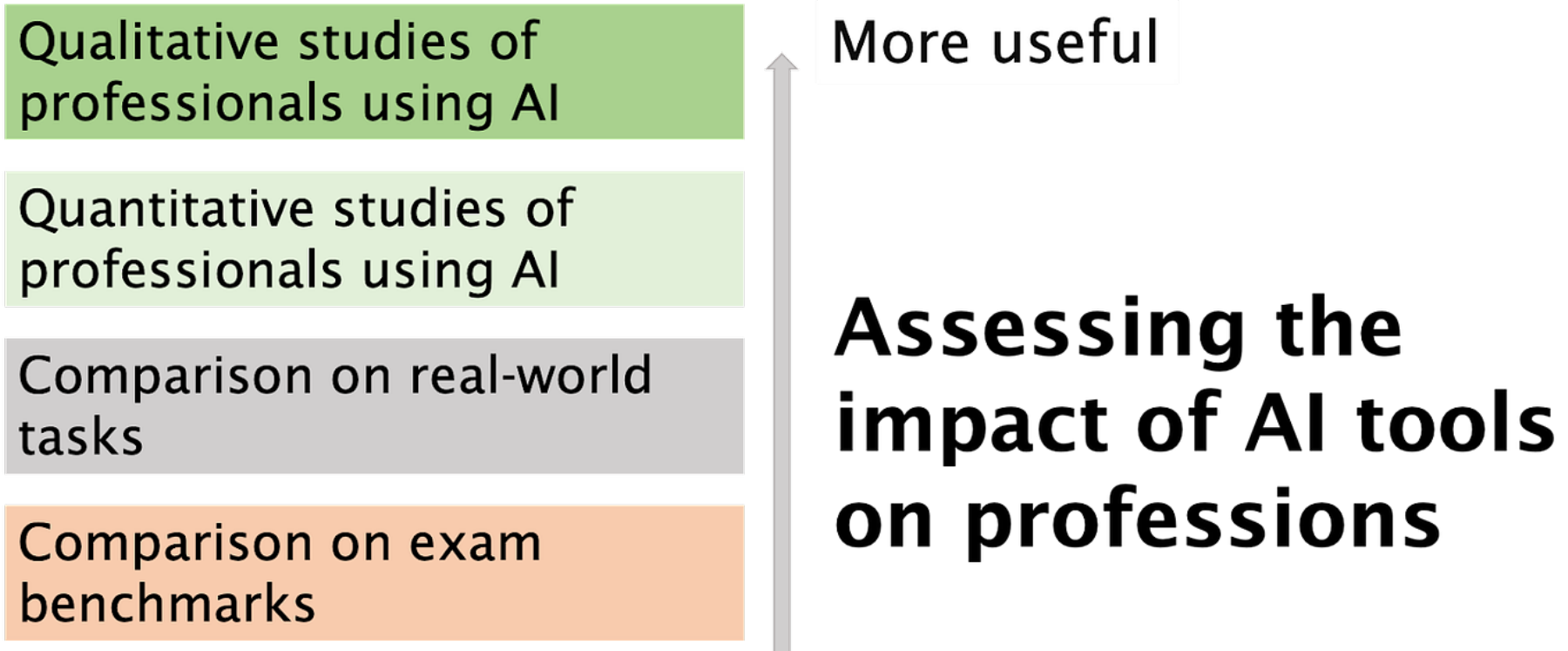
ARVIND NARAYANAN AND SAYASH KAPOOR **https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks**

Stanford | School of Medicine
MEDICINE

2024: *Large models aren't currently ready* for clinical tasks
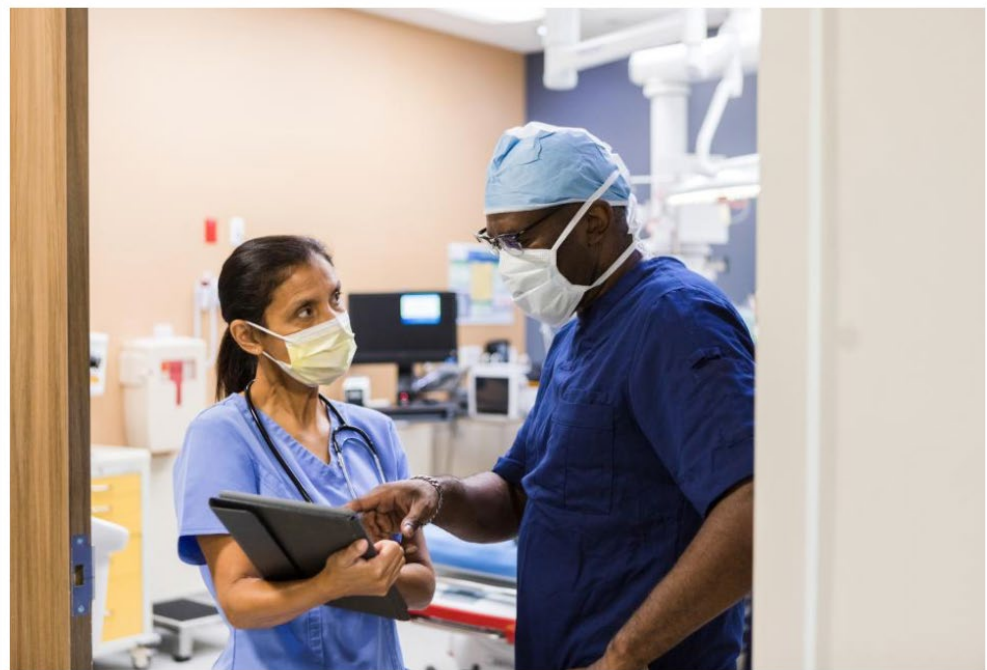
Healthcare, Language Processing, Machine Learning
How Well Do Large Language Models Support Clinician Information Needs?
Stanford experts examine the safety and accuracy of GPT-4 in serving curbside consultation needs of doctors.

Mar 31, 2023 | Dev Dash, Eric Horvitz, Nigam Shah

2023 Study @ Stanford

Twelve clinicians across multiple specialties reviewed the responses of LLMs to true to life clinical information needs. Answered two questions:

(1) "Are there elements of this answer that are so incorrect as to cause patient harm?"

(2) "The response from is similar to my assessment and would lead me to the same action"

- 91% of GPT-3.5 and 93% of GPT-4 responses are deemed safe and **the rest were considered "harmful" primarily because of the inclusion of hallucinated (fabricated) citations**

- 21% of GPT-3.5 and 41% of GPT-4 responses agreed with the known answer

2025: *Physicians aren't currently ready* for large models for clinical tasks

RCT: Large Language Model Influence on Diagnostic Reasoning

POPULATION
26 Attending physicians
24 Resident physicians

US-trained physicians with training in family medicine, internal medicine, or emergency medicine

INTERVENTION
50 Participants randomized

25 Generative artificial intelligence (AI) chatbot
Participants with access to AI chatbot were allocated 60 min to review up to 6 clinical vignettes

25 Conventional resources
Participants with access to conventional online resources such were allocated 60 min to review up to 6 clinical vignettes

FINDINGS
There was no significant difference in diagnostic performance between physicians in the chatbot group and the conventional resource group

Median diagnostic reasoning score

AI chatbot — **76%**

Conventional resources — **74%**

SETTINGS / LOCATIONS
3 Academic medical centers in the US

PRIMARY OUTCOME
Diagnostic performance, assessed using a standardized rubric including differential diagnosis accuracy, appropriateness of supporting/opposing factors, and next evaluation steps; graded via blinded expert consensus

Adjusted difference between groups:
2 percentage points (95% CI, –4 to 8 percentage points); P =.60

Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969

© AMA

Who benefits? Who does this leave out or leave behind?
How can we bridge the divide?

We need deep clinician engagement and leadership to ensure a fair, equitable and responsible tech-enabled future of dermatology care and health care

Thank you!

"Is it a beauty marks or on of a igger problem?"

Prompt: Generate a dermatology-themed New Yorker-style cartoon Imagen 3

jmko@stanford.edu

Stanford MEDICINE | School of Medicine

Working to advance the clinician-driven and patient-centered development of augmented intelligence AuI in dermatology towards meaningful outcomes

"Bytes to bedside, pixels to practice"

traind.stanford.edu